TELKOMNIKA Telecommunication Computing Electronics and Control
Vol. 20, No. 1, February 2022, pp. 98~108
ISSN: 1693-6930, DOI: 10.12928/TELKOMNIKA.v20i1.18938

Agriculture data visualization and analysis using data mining techniques: application of unsupervised machine learning

Kunal Badapanda[1], Debani Prasad Mishra[1], Surender Reddy Salkuti[2]
[1]Department of Electrical and Electronics Engineering, IIIT Bhubaneswar, Odisha, India
[2]Department of Railroad and Electrical Engineering, Woosong University, Daejeon, Republic of Korea

Article Info

Article history:

Received Dec 06, 2021
Revised Dec 12, 2021
Accepted Dec 20, 2021

Keywords:

Big data
Distplot
Elbow method
Kernel density estimate
K-means
Principal component analysis

ABSTRACT

Unsupervised machine learning is one of the accepted platforms for applying a broad data analytics challenge that involves the way to identify secret trends, unexplained associations, and other significant data from a wide dispersed dataset. The precise yield estimate for the various crops involved in the planning is a critical problem for agricultural planning. To achieve realistic and effective solutions to this problem, data mining techniques are an essential approach. Applying distplot combined with kernel density estimate (KDE) in this paper to visualize the probability density of disseminated datasets of vast crop deals for crop planning. This paper focuses on analyzing and segmenting agricultural data and determining optimal parameters to maximize crop yield using data mining techniques such as K-means clustering and principal component analysis (PCA).

Corresponding Author:

Surender Reddy Salkuti
Department of Railroad and Electrical Engineering, Woosong University
17-2, Jayang-Dong, Dong-Gu, Daejeon-34606, Republic of Korea
Email: surender@wsu.ac.kr

1. INTRODUCTION

India's agricultural history goes back to the Indus Valley Civilization. Agriculture and other related operations in India contribute (17-18)% to the Gross Domestic Product, which has a significant effect on the Indian economy. Agriculture plays an important part in India's social and economic system and is the largest economic segment in terms of demographics [1]. Crop output prediction can help the government build crop insurance policies and supply chain operation policies using big data analysis [2]. It can also help farmers by supplying them with a prediction of the past crop yield record that decreases risk management [3].

The sum of data is rising exponentially, while the speed of estimation is slowing down. Instances of large data include crop production, the field used, and crop yield. Since the government systematically and continuously gathers data on crop production and yield, the scale of the dataset is known to be big data, which is real-world data that is very difficult to interpret [4]. Statistical methods and data mining can be extended under distributed and parallel computing platforms to analyze big data and often consumes huge processing time and volume of storage to accommodate vast data sets [5]. Data mining technique plays a crucial role in data analysis. Data mining is a subfield of interdisciplinary computer science and analytics with an overall target of identifying trends, patterns, and associations within broad data sets that include strategies at the intersection of machine learning, database systems, and statistics [6]. Data mining utilizes specialized statistical algorithms with the ultimate purpose of collecting data by segmenting the data and converting the information into an understandable framework to determine the possibility of future events [7]. There are two kinds of learning approaches to data mining: unsupervised (clustering) and supervised (classifications) [8]. Clustering is the practice of evaluating a list of "data points" and sorting them according

to a distance calculation into separate "clusters" [9]. When grouping these data points, the goal should be for data points in the same cluster to be a small distance from each other, whereas data points in separate clusters should be long-distance from each other [10]. Data is grouped into well-formed classes through cluster analysis. The normal data structure can be captured by well-formed clusters [11].

This paper aims to lessen the manual work of applying data mining algorithms by using different python modules. This paper uses python-based libraries (numpy, pandas, seaborn, K-means, principal component analysis (PCA), tools, functions, and methods to quickly analyze, mine, and visualize the agriculture dataset. The dataset is visualized using distplot combined with a kernel density estimate (KDE) plot. K-means clustering technique is used in the current work to form clusters from the agricultural dataset. Compared to other clustering algorithms, the K-means algorithm is extremely simple to implement and is also very effective in computation, which may explain its popularity. The clusters obtained are visualized by reducing their dimensions using principal component analysis. The remainder of this paper is organized as follows: section 2 explains the methodology for visualizing and clustering the dataset. Section 3 presents the results and finally, section 4 concludes with some directions for future work.

2. RESEARCH METHODOLOGY

This paper aims to propose a method to analyze agricultural data using data mining techniques. Agriculture data has been obtained from credible sources in the proposed work. Input dataset consist of data with following parameters namely: crop name, production (2006-2011), area (2006-2011), yield (2006-2011) [12]. In the proposed work, the K-means clustering method is used to cluster data based on crops with identical output, area, and yield amounts [13]. Distplot combined with Kernel density estimation (KDE) is used for visualizing the probability density at different values in a continuous variable of the dataset which can improve its prediction accuracy. The principal component analysis is used for dimensionality reduction of the dataset at keeping the original information unchanged [14]. The optimum parameters for maximum output can be obtained based on this analysis.

Clustering is the process of dividing a dataset into groups such that entities in each cluster are comparatively more similar to entities of that cluster than those of the other clusters. In a dataset, Clustering can reveal undetected connections. In the proposed work, we have used the K-means algorithm to cluster our agricultural data. The K-means algorithm belongs to the prototype-based clustering group. Prototype-based methods seek to define the data set to be categorized or clustered by a (usually small) set of prototypes, particularly point prototypes, which are simply data space points [15]. Each prototype is intended to capture the distribution of a group of data points based on a definition of similarity to the prototype or closeness to its position that may be affected by the size and shape parameters of the (prototype-specific) [16]. Our goal is to group the dataset based on their similarity in characteristics, which can be accomplished using the algorithm K-means that can be summarised in the following six steps [17] in Figure 1.

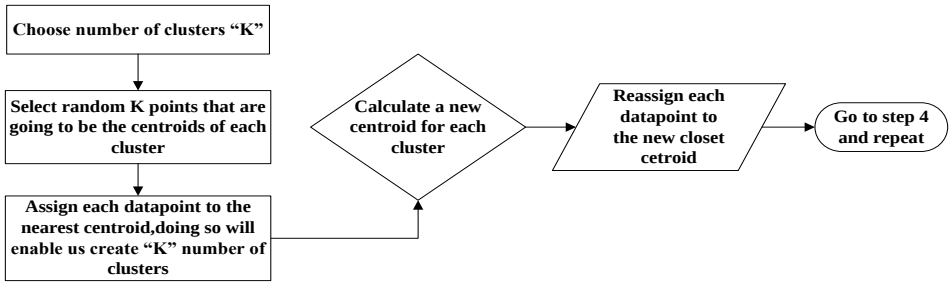

Figure 1. Steps for applying K-means clustering

Measuring similarity between objects: similarity is defined as the opposite distance, and the squared Euclidean distance between two points p and q in m-dimensional space is a commonly used distance for clustering samples with continuous features [18].

$$d(p,q)^2 - \sum_{i=1}^{n}(p_i - q_i)^2 = \|p - q\|_2^2 \quad (1)$$

Note that the index i in the preceding equation refers to the i^{th} (feature column) dimension of sample points p and q. The K-means algorithm can be defined as a simple optimization problem based on this Euclidean distance metric, an iterative approach to minimizing the sum of squares within the cluster (WCSS) [19], which is often also called cluster inertia.

$$WCSS = \sum_{i=1}^{n}\sum_{j=1}^{k} \omega^{(i,j)} \left\| p^{(i)} - \mu^{(j)} \right\|_2^2 \quad (2)$$

where $\mu^{(j)}$ is the centroid for cluster j, $\omega^{(i,j)}$ is equal to 1 if the sample $p^{(i)}$ is in cluster j, otherwise, its value is equal to 0. One of the disadvantages of this clustering algorithm is that the number of clusters, k, a priori, must be specified. Poor clustering performance may result in an inappropriate option for k. For any unsupervised algorithm, the calculation of the optimal number of clusters into which the data may be clustered is a fundamental step. One of the most common methods for evaluating this optimum k value is the elbow method [20]. Using the K-means clustering method using the sklearn python library, we are now demonstrating the provided method.

2.1. Creating and visualizing the data

Data visualization is the representation of the data values in a pictorial format. Visualization of data helps in attaining a better understanding and helps draw out perfect conclusions from the data. Data visualization plays a crucial role in any data analysis [21]. It helps to recognize which variables are important and which variables can influence our prediction model. While preparing any machine learning (ML) model we have to initially discover which characteristics are significant and how they can affect the result. This can be done by analyzing the data through data visualization.

- Python seaborn module: The data visualization modules present in Python depends on the Python Matplotlib library. Python seaborn is also one of those data visualization modules which provide functions with better efficiency and plotting features. With seaborn, data can be presented with different visualizations and different features can be added to it to enhance the pictorial representation [22].
- Distplot: A distplot or plot of distribution demonstrates the variance in the distribution of data. The Seaborn distplot can also be clubbed along with the kernel density estimate (KDE) plot to estimate the probability of distribution of continuous variables across various data values.
- KDE plot: It is a plot that depicts the probability density function of the continuous or non-parametric data variables, i.e., we can plot for the univariate or multiple variables altogether [23].
- Heatmaps: One of the important built-in functions in the direction of data exploration and visualization in seaborn is heatmaps. Seaborn heatmaps visualize the data and represent it in the form of a summary through the graph/colored maps [24]. Distplot combines two plots. It combines matplotlib. Hist function with seaborn deplot(). We have used heatmap for finding correlations in the dataset. Figure 2 describes the code for creating and visualizing the dataset, which includes 4 blocks representing the code for importing the libraries, loading the dataset, plotting the distplot, and plotting the heatmap respectively.

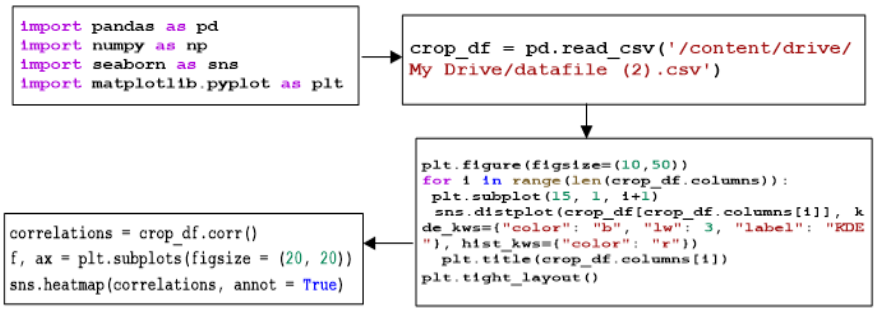

Figure 2. Steps involved in visualizing the dataset

2.2. Finding number of clusters K by elbow method

This is perhaps the best-known means of estimating the optimum number of clusters [25]. In its method, it is also a bit naive. Measure the within-clusters-sum of squares (WCSS) for various k values, and pick the k for which WCSS begins to diminish first. This is evident as an elbow in the plot of WCSS-versus-k. Within-cluster-sum of Squares sounds sort of complicated. Let's break down this in Figure 3.

We need to scale the continuous features to give all characteristics equal significance. Scikit-learn's standard scaler will be included. We will initialize K-means for each k value and use the attribute of inertia to define the number of squared sample distances to the nearest cluster core. The squared distance number tends to zero as k increases. Imagine that k is set to its maximal value n (where n is the number of samples) and each sample forms its cluster, meaning the total of square distances is equal to zero. The code used to map the total k square distances is defined in Figure 4. This figure depicts the four blocks representing the code

for importing the libraries, scaling the dataset, initializing the K-means for each k value, and applying the elbow method, respectively. If the plot looks like an arm, so an ideal k is the elbow on the arm. Using the sklearn library and our feature for calculating WCSS for several values for k, let us implement this in Python.

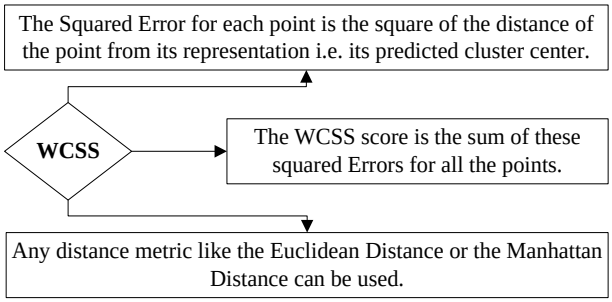

Figure 3. Brief description of WCSS

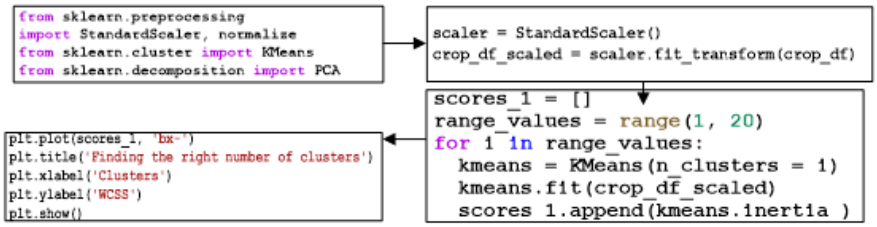

Figure 4. Steps involved in finding the number of clusters (K) by elbow method

2.3. Applying K-means and principal component analysis (PCA)

In the code for applying the K-means algorithm, the K-means object has been created and passed as the number of clusters "K" obtained from the elbow method. In the next line fit method on K-means has been called and the "crop_df_scaled" dataset has been passed through it K-means. Labels_ is used to see the labels for the datapoints. Via dimensionality reduction, the clusters we have identified after applying the K-means clustering approach can be visible. PCA is an effective tool for visualizing high-dimensional data in combination with K-means. It is an unsupervised machine learning algorithm. PCA projects them into a lower-dimensional vacuum, restricts them, and visualizes them to only a few significant key ones [26]. Figure 5 describes the code for implementing PCA on the dataset, each block in this figure represents the code for obtaining the principal components, creating a dataframe with two components, c concatenating the labels to the dataframe, and visualizing and interpreting the clusters.

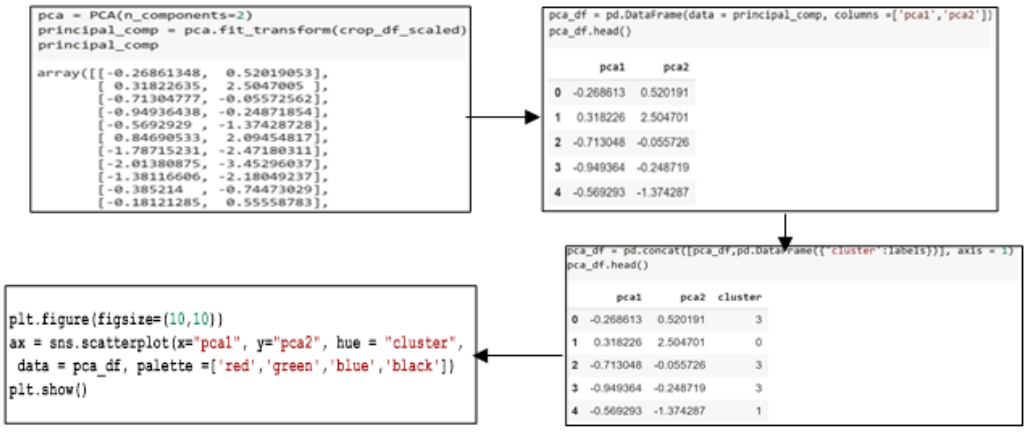

Figure 5. Steps involved in applying PCA on the dataset

3. RESULTS AND DISCUSSION

3.1. Visualizing the dataset

The dataset must be visualized before applying the K-means algorithm to the dataset. Results of data visualization are shown in Figures 6 (see Appendix) and 7. Figure 6 (see Appendix) depicts the KDE plot combined with distplot is plotted for the dataset to analyze the data through visualization. Figure 7 depicts the result of the heatmap plot which is plotted by representing the dataset in the form of a 2-dimensional format for finding correlations among the data.

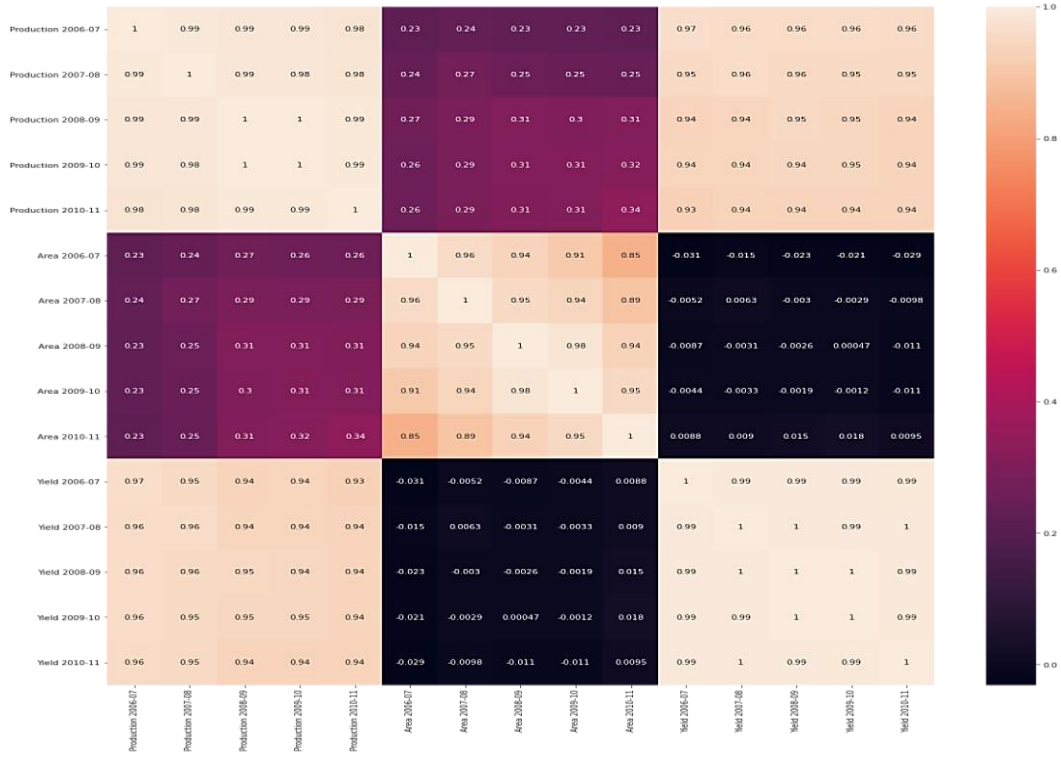

Figure 7. Heatmap for the given dataset

3.2. Clustering

To calculate the K value (number of clusters), the elbow method is applied to the dataset. The outcome of the elbow process is represented in Figure 8, and it depicts the result of the elbow plot which is plotted using the within-cluster sum of squares for a range of values of K. The optimum number of clusters (K value) is determined by choosing the "elbow" value of K, i.e., the point at which the WCSS starts to decrease linearly. Therefore, we assume that the number of clusters is 4 for the given dataset. Table 1 depicts the result of the K-means clustering algorithm. Figure 9 depicts the clusters we have obtained, represented by reducing their dimensions using Principal component analysis. Crops are commonly picked for their economic significance. The agricultural planning process, however, involves an estimate of the yield of many crops. In this context, using data availability as the main metric, 54 crops have been selected for this work. Crops were only chosen when appropriate data samples came under review in the 6-year range (2006-11).

As a result of the K-means clustering algorithm, 4 clusters are formed. Cluster 0 represents the crops having medium production, high area, and medium-low yield. Cluster 1 represents the crops having low production, low area, and medium yield. Cluster 2 represents the crops having high production, medium area, and high yield. Cluster 3 represents the crops having medium-low production, medium-low area, and low yield. Principal component analysis is used to represent the clusters by reducing their dimensions. The present work covers the distplot combined with the kernel density estimate plot and heatmap for visualization. The elbow method is used for finding the optimal number of clusters "K". K-means clustering algorithm is applied to form clusters from the dataset. The principal component analysis is used to represent the clusters formed by reducing their dimensions. The crop data collection can be analysed using these methods and the optimum parameters for crop production can be calculated.

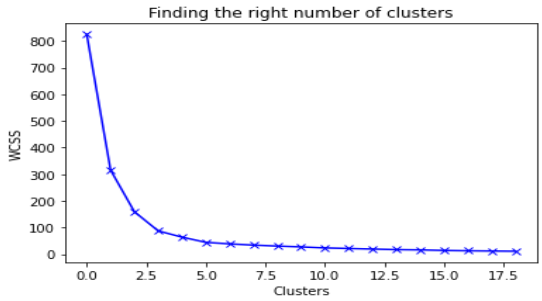

Figure 8. WCSS vs K plot (elbow method)

Table 1. Clusters obtained from the K-mean algorithm to represent crops as per production, area, and yield

Cluster	Crops	Production range	Area range	Yield range
Cluster 0 (Medium Production, High Area and Medium-Low Yield)	Rice, Maize, Soyabean, Dry ginger, Arecanut, Garlic, Total Fruits & Vegetables, Potato, Onion, Banana	199.59-299.95	168.56-213.63	119.57-140.70
Cluster 1 (Low Production, Low Area, and MediumYield)	Bajra, Ragi, Small millets, Barley, Sesamum, Rapeseed & Mustard, Linseed, Safflower, Niger seed, Mesta, Jute & Mesta, Sannhamp, Dry chilies, Cardamom, Coriander, Sweet potato, Tobacco	97.27-120.54	71.24-76.68	134.70-154.15
Cluster 2 (High Production, Medium Area, and High Yield)	Total Spices	1427.70-1790.60	121.30-136.60	1172.10-1310.80
Cluster 3 (Medium-Low Production, Medium-Low Area, and Low Yield)	Total Foodgrains, Wheat, Jowar, Coarse Cereals, Cereals, Gram, Arhar, Other Pulse, Total Non-Food grains, Total Oilseeds, Groundnut, Castor seed, Sunflower, Nine Oilseeds, Coconut, Cottonseed, Total Fibers, Cotton (lint), Jute, Tea, Coffee, Rubber, Black pepper, Turmeric, Tapioca, Sugarcane.	150.048-174.83	121.25-126.68	123.97-137.51

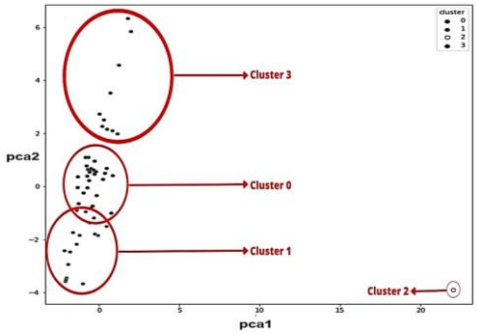

Figure 9. Result of PCA

4. CONCLUSIONS AND FUTURE WORK

In developing countries such as India, agriculture is the most significant application field. In agriculture, the use of information technologies can improve the decision-making scenario, and farmers can perform more. In several matters relating to the agriculture sector, data mining plays a key role in decision-making. This paper discusses the role of data mining from the perspective of the agriculture field. On the input data, different data mining techniques are applied that can be used to determine the best output yielding process. To obtain the optimum parameters to achieve higher crop yield, the present study used data mining techniques such as K-means clustering, principal component analysis. Through this paper, an attempt is made to lessen the manual work of applying data mining algorithms by using different python modules. Expanding the present work to evaluate soil, climate conditions, demand data, and other variables for the crop to improve the crop yield is scope for future work.

ACKNOWLEDGEMENTS

This research work was funded by "Woosong University's Academic Research Funding - 2022".

APPENDIX

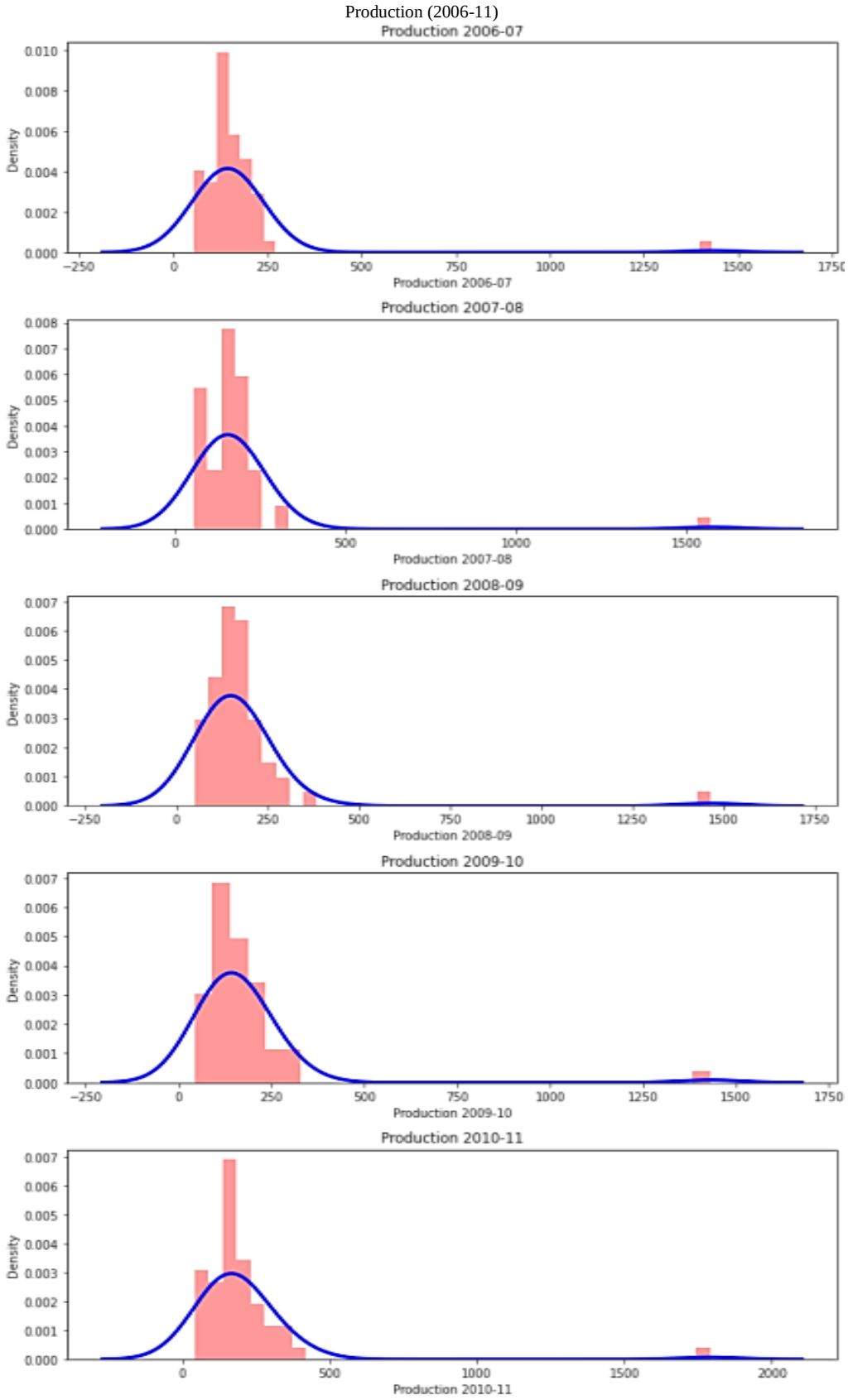

Figure 6. Distplot combined with KDE plot for the given dataset

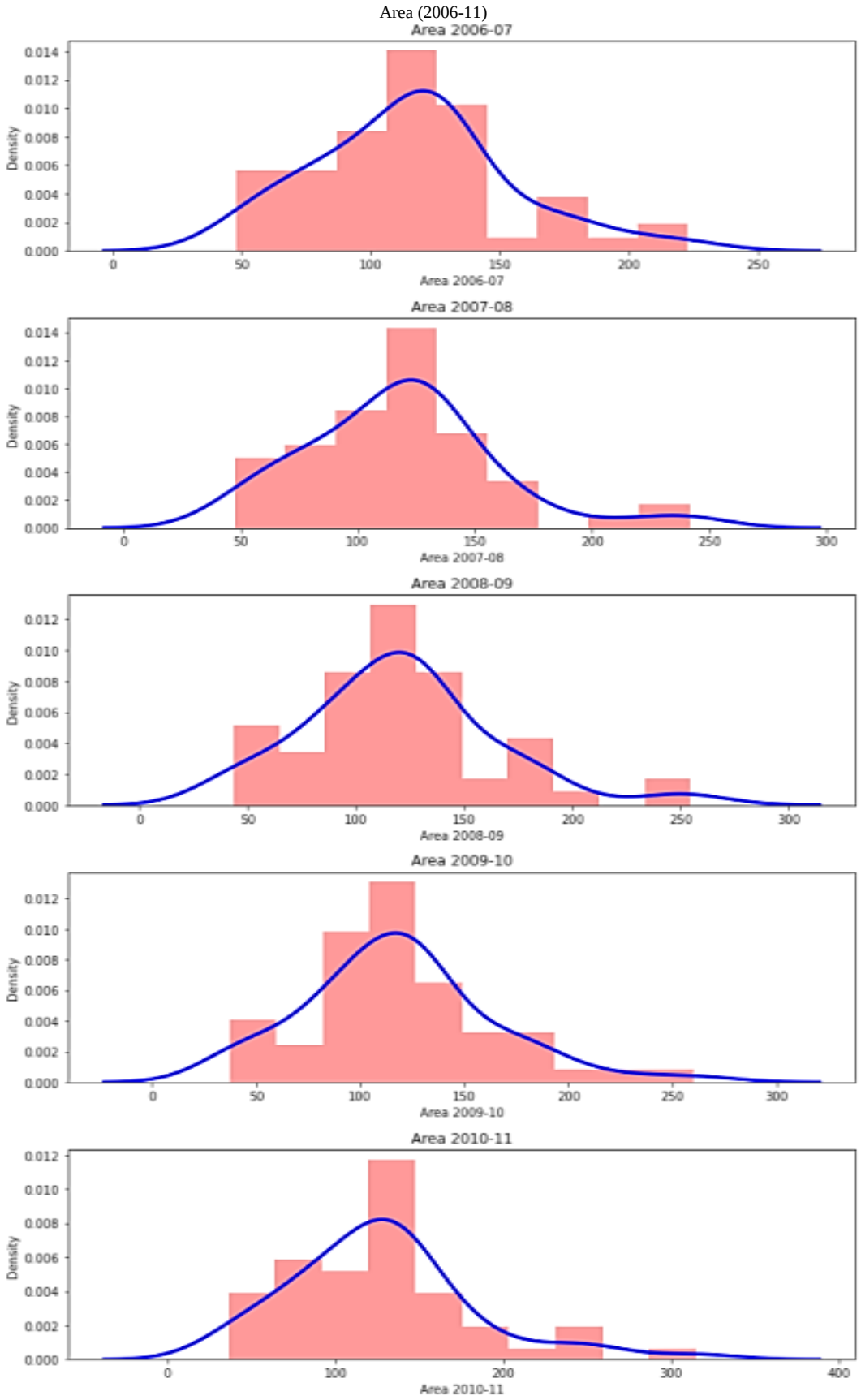

Figure 6. Distplot combined with KDE plot for the given dataset (continue)

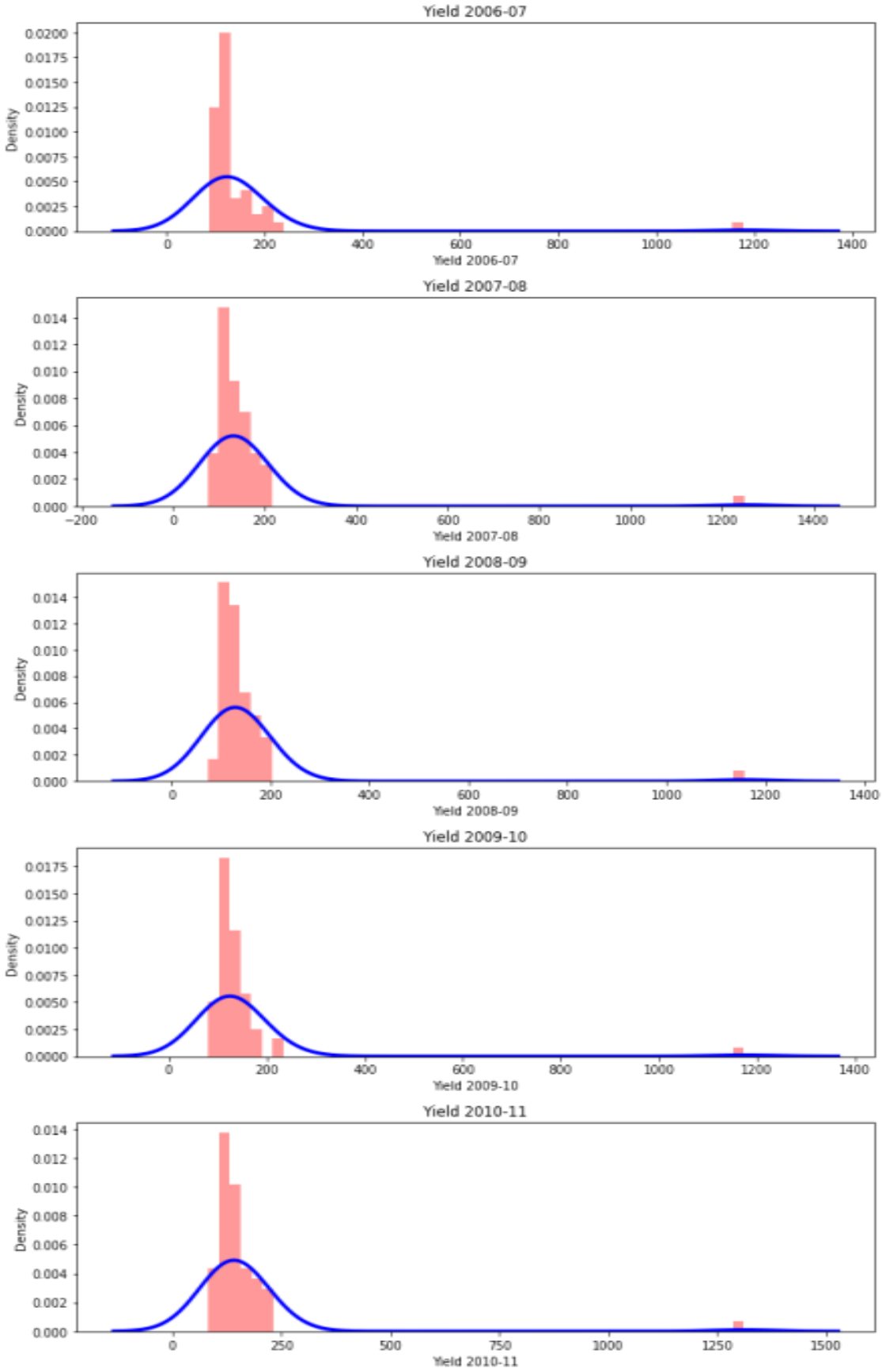

Figure 6. Distplot combined with KDE plot for the given dataset (continue)

REFERENCES

[1] R. R. Wagh and A. Dongre, "Agricultural sector: status, challenges and it's role in Indian economy," *Journal of Commerce and Management Thought*, vol. 7, no. 2, pp.209-218, 2016, doi: 10.5958/0976-478X.2016.00014.8.

[2] A. K. S., Md. Khan Tajuddin, and K. Avinash, "Adoption of crop insurance and impact: insights from India," *Agricultural Economics Research Review*, vol. 31, no. 2, pp.163-174, 2018, doi: 10.5958/0974-0279.2018.00034.4.

[3] K. Coble, T. O Knight, G. F. Patrick, and A. E. Baquet, "Crop producer risk management survey: a preliminary summary of selected data," Mississippi State University, Information Report 99-001, Mississippi, 1999.

[4] X. Wu, X. Zhu, G. Q. Wu, and W. Ding, "Data mining with big data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 01, pp. 97-107, 2014, doi: 10.1109/TKDE.2013.109.

[5] C. Wang, M.-H. Chen, E. Schifano, J. Wu, and J. Yan, "Statistical methods and computing for big data," *Stat Interface*, vol. 9, no. 4, pp. 399-414, 2016, doi: 10.4310/SII.2016.v9.n4.a1.

[6] S. Agarwal, "Data mining concepts and techniques," *International Conference on Machine Intelligence Research and Advancement*, 2013, pp. 203-207, doi: 10.1109/icmira.2013.45.

[7] M. Shafiei and E. Milos, "A statistical model for topic segmentation and clustering," *Conference of the Canadian Society for Computational Studies of Intelligence*, Berlin, 2008, doi: 10.1007/978-3-540-68825-9_27.

[8] K. Bindra and A. Mishra, "A detailed study of clustering algorithms," *6th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, Noida, India, 2017, pp. 370-376, doi: 10.1109/ICRITO.2017.8342454.

[9] C. Maionea, D. R..Nelson, and R. M. BarboSsa, "Research on social data by means of cluster analysis," *Applied Computing and Informatics*, vol. 15, no. 2, pp. 153-162, 2019, doi: 10.1016/j.aci.2018.02.003.

[10] A. Khandare and A. Alvi, "Efficient clustering algorithm with improved clusters quality," *IOSR Journal of Computer Engineering (IOSR-JCE)*, vol. 18, no. 6, pp. 15-19, 2016, doi: 10.9790/0661-1806051519.

[11] M. Z. Rodriguez *et al*., "Clustering algorithms: A comparative approach," *PLoS ONE*, vol. 14, no. 1, pp. 02-36, 2019, doi: 0.1371/journal.pone.0210236.

[12] J. Majumdar and S. Naraseeyappa, "Analysis of agriculture data using data mining techniques: application of big data," *Journal of Big Data*, vol. 4, no. 1, 2017, doi: 10.1186/s40537-017-0077-4.

[13] T. K. Anderson, "Kernel density estimation and K-means clustering to profile road," *Accident Analysis & Prevention*, vol. 41, no. 3, pp. 359-364, 2009, doi: 10.1016/j.aap.2008.12.014.

[14] S. Sehgal, H. Singh, M. Agarwal, and V. B., Shantanu, "Data analysis using principal component analysis," *International Conference on Medical Imaging, m-Health and Emerging Communication Systems (MedCom)*, 2014, pp. 45-48, doi: 10.1109/MedCom.2014.7005973.

[15] J. Kim, R. Krishnapuram, and R. Dave, "Application of the least trimmed squares technique to prototype-based clustering," *Pattern Recognition Letters*, vol. 17, no. 6, pp. 633-641, 1996, doi: 10.1016/0167-8655(96)00028-1.

[16] J. Hämäläinen, S. Jauhiainen, and T. Karkkainen, "Comparison of Internal Clustering Validation Indices for Prototype-Based Clustering," *Clustering Algorithms*, vol. 10, no. 3, pp. 105-119, 2017, doi: 10.3390/a10030105.

[17] S. Na, L. Xumin, and G. Yong, "Research on k-means clustering algorithm: an improved k-means clustering algorithm," *Third International Symposium on Intelligent Information Technology and Security Informatics*, Jinggangshan, China, 2010, pp. 63-67, doi: 10.1109/IITSI.2010.74.

[18] J. Irani and N. Pise, and M. Phatak, "Clustering techniques and the similarity measures used in clustering: A survey," *International Journal of Computer Applications*, vol. 134, no. 7, pp. 9-14, 2016, doi: 10.5120/ijca2016907841.

[19] T. Thinsungnoena, N. Kaoungku, P. Durongdumronchaib, K. Kerdprasop, and N. Kerdprasop, "The clustering validity with silhouette and sum of squared errors," *3rd International Conference on Industrial Application Engineering*, 2015, doi: 10.12792/ICIAE2015.012.

[20] D. Marutho, S. H. Handaka, E. Wijaya, and Muljono, "The determination of cluster number at k-mean using elbow method and purity evaluation on headline news," *International Seminar on Application for Technology of Information and Communication*, Semarang, Indonesia, 2018, doi: 10.1109/ISEMANTIC.2018.8549751.

[21] S. K. A. Fahad and A. E. Yahya "Big data visualization: allotting by R and Python with GUI Tools," *International Conference on Smart Computing and Electronic Enterprise*, Kuala Lumpur, Malaysia, 2018, doi: 10.1109/ICSCEE.2018.8538413.

[22] P. Lemenkova, "Python libraries matplotlib, seaborn and pandas for visualization geospatial," *Analele stiintifice ale Universitatii Alexandru Ioan Cuza" din Iasi - seria Geografie*, vol. 64, no. 1, pp. 13-32, 2020, doi: 10.15551/scigeo.v64i1.386.

[23] T. Ledl, "Kernel density estimation: theory and application in discriminant analysis," *Austrian Journal of Statistics*, vol. 33, no. 3, pp. 267-279, 2004, doi: 10.17713/ajs.v33i3.441 .

[24] S. Zhao, Y. Guo, Q. Sheng, and Y. Shyr, "Advanced heatmap and clustering analysis using heatmap3," *BioMed Research Internationa*, vol. 2014, pp. 1-6, 2014,doi: 10.1155/2014/986048.

[25] M. A. Syakur, B. K. Khotimah, E. M. S. Rochman, and B. D. Satoto, "Integration K-means clustering method and elbow method for identification of the best customer profile cluster," *IOP Conference Series: Materials Science and Engineering, vol. 336, The 2nd International Conference on Vocational Education and Electrical Engineering (ICVEE), Surabaya, Indonesia*, Nov. 2017, doi: 10.1088/1757-899X/336/1/012017.

[26] L. L. Palese, "A random version of principal component analysis in data clustering," *Computational Biology and Chemistry*, vol. 73, pp. 57-64, 2018, doi: 10.1016/j.compbiolchem.2018.01.009.

BIOGRAPHIES OF AUTHORS

Kunal Badapanda received the Bachelor of Technology (B.Tech) in the stream of Electrical and Electronics Engineering in International Institute of Information Technology Bhubaneswar (IIIT-BH), Odisha, India. His research interests include Power Electronics, Artificial Intelligence, Data Mining and Machine Learning. He can be contacted at email: b317052@iiit-bh.ac.in.

Debani Prasad Mishra received the B.Tech. in electrical engineering from the Biju Patnaik University of Technology, Odisha, India, in 2006 and the M.Tech in power systems from IIT, Delhi, India in 2010. He has been awarded the Ph.D. degree in power systems from Veer Surendra Sai University of Technology, Odisha, India, in 2019. He is currently serving as Assistant Professor in the Dept of Electrical Engg, International Institute of Information Technology Bhubaneswar, Odisha. He has 11 years of teaching experience and 2 years of industry experience in the thermal power plant. He is the author of more than 80 research articles. His research interests include soft Computing techniques application in power system, signal processing and power quality. 3 students have been awarded Ph.D under his guidance and currently 4 Ph.D. Scholars are continuing under him. He can be contacted at email: debani@iiit-bh.ac.in.

Surender Reddy Salkuti received the Ph.D. degree in electrical engineering from the Indian Institute of Technology, New Delhi, India, in 2013. He was a Postdoctoral Researcher with Howard University, Washington, DC, USA, from 2013 to 2014. He is currently an Associate Professor with the Department of Railroad and Electrical Engineering, Woosong University, Daejeon, South Korea. His current research interests include power system restructuring issues, ancillary service pricing, real and reactive power pricing, congestion management, and market clearing, including renewable energy sources, demand response, smart grid development with integration of wind and solar photovoltaic energy sources, artificial intelligence applications in power systems, and power system analysis and optimization. He can be contacted at email: surender@wsu.ac.kr.